Potential of an educational application of a language statistical translation system

Aji P. Wibawa†, Andrew Nafalski‡ & Zorica Nedic‡

State University of Malang, Malang, Indonesia†
University of South Australia, Adelaide, Australia‡

ABSTRACT: This article presents the potential application of a statistical machine translation system to support Indonesian language teaching at school. Advanced Javanese-to-Indonesian Statistical Machine Translation (AJI-SMT) is used as a case study. This statistical machine translation provides 20 types of Javanese-Indonesian language and speech levels translations. As a result, the use of AJI-SMT in a communicative approach is possible. In a class activity, students may use AJI-SMT to translate the dialogue in a role-play scenario. A proper translation will maintain the socio-linguistic aspects of the Javanese speech levels that are related to the use of various politeness levels based on the characteristics of the interlocutors, depending on their social status and the seniority levels. The translation is not an easy task as each of the Javanese language politeness level constitutes a separate language bound only by the grammatical structure but in most cases not a common vocabulary. AJI-SMT has the potential to contribute to saving the cultural heritage of the Javanese languages.

INTRODUCTION

The basic requirement for understanding of a particular subject at school is to know the meaning of each word within the subject. Without knowing the words, knowledge achievement is impossible. That knowledge would be easier to understand if it were stated in the learner's first language (L1). When knowledge is delivered in a second language (L2), a translator, either human or machine, is probably needed.

Computer scientists have been developing machine translation (MT) to translate text or speech between two or more natural languages since the early 1950s. The basic idea of translation software is to transform a word from one language into a word in a different language. However, this simplistic approach frequently results in an inaccurate translation. Therefore, research into MT is growing rapidly in order to improve the translation performance.

Statistical machine translation (SMT) approach is popular and rapidly developing. Google Translate and Bing Translator are examples of publicly available statistical machine translators [1]. This article explores the potential application of SMT system to support learning process at school. The first discussion concerns the state of the art of SMT technology. Later, this article explains a specific SMT system, called Advanced Javanese-to-Indonesian Statistical Machine Translation (AJI-SMT), that has been used here as a case study. The final discussion focuses on the use of AJI-SMT to support Javanese language learning at school.

STATISTICAL MACHINE TRANSLATION

Statistical machine translation relies on statistical modelling while performing bilingual translations. SMT is also called probabilistic machine translation, since it relies heavily on probability calculations [2]. The trend has been to go away from word-base translation towards phrase-based approach since simplistic word-to-word MT is powerless to translate collocations. More specifically, recent research on SMT focuses more on translation of different length sentences using phrase-based approaches [3][4], syntax-based SMT [3][5][6] and hierarchical phrase-based SMT [3]. While the syntax-based SMT focuses on translation of syntactic units, hierarchical models divide phrases into smaller translation units as well as applying context-free grammars as translation rules.

Researchers may develop their SMT using specific translation models such as the IBM models [7], source-channel models [8] and Bayes rule decision models [9]. Furthermore, widely used development tools, such as Giza++ [6][10][11] and Moses [12][13] facilitate researchers in their development of SMT based on these models. However, using these particular models as the foundation of SMT development is optional. Simple probability approaches to complex statistical inference methods can be used to obtain optimal decision making in SMT [2]. Therefore, the

researcher or SMT developer can use or improve the *ready to use* models or alternatively build the SMT model from scratch [14].

The statistical approach is effective in translating any language unless the corpora of the target language are insufficient or unavailable. However, the explicit use of statistical inference is better for dealing with translational divergence, which is a frequent problem with example based machine translation (EBMT) [15][16]. Furthermore, the SMT is mathematically simpler and developed faster than rule based models [9]. Bearing those advantages in mind, statistical machine translation is selected as the core system in this research.

ADVANCED JAVANESE-TO-INDONESIAN STATISTICAL MACHINE TRANSLATION

The Javanese language is the most widely used local language of Indonesia, and is used by approximately 75 million speakers [17][18]. While the number of its users can be an indication that the language is far from being endangered, some studies found that most Javanese youngsters cannot speak respectfully using this refined language [19][20].

The emotional maturity of a Javanese person can be judged from the way he or she communicates using various speech levels. A Javanese person is not properly addressed as an adult unless he or she is able to use the polite form (*krama*) skilfully [19]. Furthermore, the use of varying polite forms indicates the speaker's linguistic knowledge, as well as pragmatic sensitivity and refinement [21].

In fact, speaking and writing of the refined form (*krama*) are increasingly less evident in Javanese society. Today, only people older than 70 years of age, shadow puppeteers (*dhalang*), ritual specialists and traditional dancers, are highly competent in using Javanese speech levels [22]. As a result, mistakes are commonplace and found more frequently when the language is used. Since speech levels capture a great deal of the Javanese cultural context, improper use of these politeness forms is not only dangerous to the language existence but is also harmful to the Javanese culture.

AJI-SMT was created in reaction to this threatening situation. The MT translates entered text, corresponding to what a user might say, into a proper speech level.

Javanese Speech Levels

The speech levels are defined as use of specific words to communicate with a particular person [23][24].These levels were initiated in 1626 [23]. Since then, the use of speech levels has been inseparable from the Javanese culture. Speech level is an indicator of psychological maturity [19] and the rule of polite communication [23].

The Javanese Congress in 1991 simplified the speech levels into four categories: *ngoko* (Ng), *ngoko alus* (NgA), *krama* (Kr) and *krama aIus (KrA).* AJI-SMT was created based on this classification. The first speech level *ngoko* is basically used to speak with familiar, younger or low social status people. For example, parents talk to their children; a teacher explains the subject to his/her students; a brother speaks to his little sister; a master commands his maid or someone talks to himself.

Ngoko alus, refined *ngoko*, is applied when someone interacts with either a younger or older respected person. The third classification is *krama*, mainly used to speak with friends with whom one is not familiar, unfamiliar younger people and strangers. Finally, *krama alus* (*krama inggil*) is the most formal and refined speech. People should use this level when they converse with their parents, teachers, superiors, masters and other older persons. For instance, there is a situation where a student asks his/her teacher about something. The student should apply *krama inggil* to ask politely and, then, his/her teacher will answer him using *ngoko*.

Basically, *ngoko* and *krama* levels use *ngoko* and *krama* words in conjunction with their affixes. Example (1) and (2) show the affix transformation in various Javanese speech levels. *Krama inggil* is combined with the corresponding vocabularies in both *ngoko alus* and *krama alus* levels. Furthermore, different vocabularies are used in most politeness levels in order to express Javanese pronouns.

1)	Aku	diceluk	Ibu-ku	(Ng)
	Aku	ditimbali	Ibu -ku	(NgA)
	Kula	dipun-timbali	Ibu kula	(Kr)
	Kawula	dipun-timbali	Ibu kawula	(KrA)
	I	call (passive)	mother-my	
	My mother calls me			
2)	Bapak-mu	lagi	mangan	(Ng)
	Bapak-mu	lagi	mangan	(NgA)
	Bapak-sampeyan	saweg	dhahar	(Kr)
	Bapak-panjenengan	nembe	dhahar	(KrA)
	Father-you	*progressive* marker	eat	
	Your father is eating			

Features and Performance

Users may use AJI-SMT to translate *bahasa Indonesia* to any Javanese speech level, from any Javanese speech level to *bahasa Indonesia*, and between any two Javanese speech levels. The advanced features of AJI-SMT are the source language classifier (SLC) and target language selector (TLS) used to assist the user in classifying their words before translating to a proper speech level based on the characteristics of their interlocutor.

Figure 1 illustrates the features of AJI-SMT. In this integrated translator, user input consists of source text and the characteristics of an interlocutor. The SLC identifies the language of the source text, while the TLS chooses the variant of target language based on the interlocutor's characteristics. Subsequently, the machine translation is executed when the detected source language and target language (derived from the language selection rules) are at different levels. These decision support systems help inexperienced users to decide whether it is necessary to translate their phrases or not.

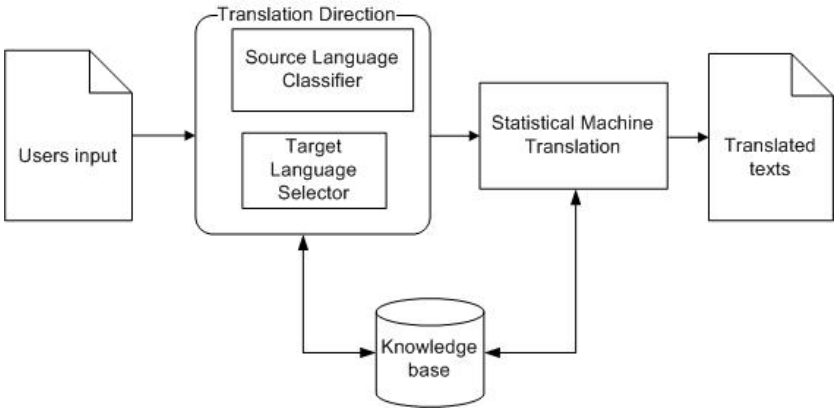

Figure 1: Advanced Javanese-to-Indonesian statistical machine translation.

The statistical machine translation is based on an alignment algorithm that uses an edit shifting distance coefficient [25] and also impossible pair elimination [26]. All twenty bilingual combinations from the five languages under consideration were tested. As a result, the overall accuracy of this SMT shows the relatively satisfactory result that 73.8% of translations tested were accurate. Furthermore, evaluation of translation quality shows that 82.8% of the time the system generates a result that is of acceptable quality indicating that the SMT is reasonably efficient.

The causes of inaccurate translations are dominated by probabilistic errors, which can be simply mitigated by increasing the size of the parallel text used in the training phase. A Naive Bayes classifier was selected as the SLC based on overall accuracy (accurate classification 88% of the time). This robust method can easily embed to the main system (SMT) since both are based on probability. The target language selector (TLS) selects the translation's target language based on user inputs of the difference in age, social status and closeness of relationship between the user and an interlocutor. The developed communication rules are simple but effective in terms of target language selection.

AJI-SMT integrates the SMT with the SLC and TLS advanced subsystems. This integrated translation system is, then, examined to measure its performance. Even though some mistakes still occur, the overall performance of this integrated system is quite satisfactory, with an average of 66% of test results being classified as accurate over all language combinations.

AJI-SMT TO SUPPORT LANGUAGE AND CULTURAL LEARNING

Javanese, including its speech levels is studied for six years at school. In fact, most youngsters experience difficulties when they try to learn the polite language because of the vapidity, complexity and scarcity of learning materials [27]. Nowadays, Communicative Language Teaching (CLT) is used on the Javanese curriculum to create more fascinating language-learning situation [28-30]. The communicative approach emphasise the use of the language in real communication situations. Interactive classroom activities, such as role-play, quizzes and games may naturally increase the students' target language motivation of learning and, thus, the learning efficiency.

In this article, role-play has been selected as the learning activity since communication using speech levels is based on the social characteristics of the interlocutor. The interlocutor should respond the speaker with different speech levels if their social statuses are unequal. On the other hand, symmetric communication using the same speech levels happens when speaker and interlocutor are of the same social status.

In role-play scenarios, students may act as different people to demonstrate both symmetric and asymmetric communication in Javanese. Accordingly, the class activity divided into three scenarios: pre-translation, translation activity and post translation will be discussed in the following sub-sections.

Pre-translation Activity

Pre-translation activity performs symmetric communication between two people with equal or different social status. The scenario is, then, detailed into the following steps.

Step 1: The teacher prepares various dialogues between two people. All of them must be in *ngoko* (the basic level) as showed in the following examples. The first example, 3) is a dialogue between Anin and his classmate Radya while the second, 4) is the dialogue between the father Pak Bowo and his son Raka.

3) Anin: *Radya kowe apa wis garap PR ?* (Radya, did you do your homework?)
Radya: *Lagi separo, PR'e angel, kowe piye?* (Still half of it, they are difficult, how about yours?)
Anin: *Padha, ayo digarap bareng bae* (is about the same, let us do it together)
4) Pak Bowo: *Le, aja rame bae, wis bengi, wayahe turu* (Son, please keep silent, it is a bedtime)
Raka: *Limang menit maneh, lagi aku mapan turu* (five minutes more, after that I will go to sleep)

Step 2: Two selected students perform a mono level dialog.

Step 3: The class discusses the content of the dialogue. Teacher asks questions to the class related to the content of the dialogue.

Step 4: Students recognise the characteristics of both first (S1) and second (S2) speakers. As seen in Table 1, if the characteristics are equal the dialogue translation is not necessary.

Table 1: Dialogue characteristics.

Dialogue	S1	S2	Comparison between S1 and S2			Type	Translation
			Age	Relationship	Social status		
3	Anin	Radya	similar	close	equal	symmetric	no
4	Pak Bowo	Raka	older	close	higher	asymmetric	yes

Translation Activity

Based on the classification shown in Table 1, only the second dialogue (4) is suitable for translation activity. The teacher divides the class into two groups. The first group (G1) translates the text manually using a Javanese dictionary. The second group (G2) uses AJI-SMT to translate the dialogue. Their activities are detailed as follows:

Step 1: Students enter the speaker's text (S1) and interlocutor (S2) characteristic. Table 2 shows the inputs of AJI-SMT.

Table 2: Inputs of AJI-SMT.

Speaker's Text	Interlocutor	Interlocutor's characteristic		
		Age	Relationship	Social status
Le, aja rame bae, wis bengi, wayahe turu	S2	younger	close	lower
Limang menit maneh, lagi aku mapan turu	S1	older	close	higher

Step 2: Students translate the inputted texts using AJI-SMT. As seen in Table 3, there is no difference between the first source text and its translation. In this case, AJI-SMT is not executed and the source text is maintained because the speaker status is higher than the interlocutor.

Table 3: Translation using AJI-SMT.

Source Text	Translation	Translation level
Le, aja rame bae, wis bengi, wayahe turu	*Le, aja rame bae, wis bengi, wayahe turu*	*ngoko* (Ng)
Limang menit maneh, lagi aku mapan turu	*Gangsal menit malih, nembe kula mapan tilem.*	*krama* (Kr)

Step 3: Students repeat Steps 1 and 2 for the speaker's text (S2) and interlocutor (S1) characteristics as seen in the last line of Table 2 and Table 3.

Post-translation Activity

The aim of post-translation activity is to evaluate the results of both human and machine translation.

Step 1: Both G1 and G2 exchange their translation results.

Step 2: Students compare and discuss the translations.

Step3: The teacher justifies the discussion by showing the best translation result.

CONCLUSIONS

The use of a statistical translation system in the educational sector is a promising development. The translation system elaborated here is based on some 20 translation pathways, each supported by a sophisticated software system that has been created and described in this article.

In the case of learning Javanese speech levels, students can use the AJI-SMT as a supportive tool of the communicative approach, particularly in a role-play activity. The translation helps students to perform both symmetric and asymmetric communication in the Javanese and Indonesian languages. As a result, the Javanese language politeness levels can be used properly and more frequently, which may preserve the regional languages from becoming extinct.

REFERENCES

1. Madnani, N., iBLEU: Interactively Debugging and Scoring Statistical Machine Translation Systems. IEEE (2011).
2. Wu, D., MT model space: statistical versus compositional versus example-based machine translation. *Machine Translation,* 19, 213-227 (2007).
3. Xiao, T., Zhu J. and Liu, T., Bagging and Boosting statistical machine translation systems. *Artificial Intelligence*, 195, 496-527 (2013).
4. Silvestre-Cerdà, J.A., Andrés-Ferrer, J. and Civera, J., Explicit length modelling for statistical machine translation. *Pattern Recognition,* 45, 3183-3192 (2012).
5. Dong, X., Xue, H. and Yang, Y., Factor-based Uyghur-Chinese statistical machine translation. *Inter. J. of Advancements in Computing Technol.*, 4, 275-283 (2012).
6. Hassan, H., Hearne, M., Way, A. and Simaan, K., Syntactic phrase-base statistical machine translation. *Proc. 2006 IEEE Spoken Language Technol. Workshop*, Palm Beach, Aruba, 238-241 (2006).
7. Deng, Y. and Byrne, W., HMM word and phrase alignment for statistical machine translation. *IEEE Transactions on Audio, Speech, and Language Processing,* 16, 494-507 (2008).
8. Macherey, K., Bender O. and Ney, H., Applications of Statistical Machine Translation Approaches to Spoken Language Understanding (2009).
9. Sripirakas, S., Weerasinghe, A.R. and Herath, D.L., Statistical machine translation of systems for Sinhala-Tamil. *Proc. 2010 Inter. Conf. on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka, 62-68 (2010).
10. Zhengxian, G. and Guodong, Z., Employing topic modeling for statistical machine translation. *Proc. 2011 IEEE Inter. Conf. on Computer Science and Automation Engng.*, Shanghai, China, 24-28 (2011).
11. Huang, C-C., Chen, W-T. and Chang, J.S., Bilingual segmenter for statistical machine translation. *Proc. 2008 Second Inter. Symp. on Universal Communication*, Osaka, Japan, 97-104 (2008).
12. Dong, X., Yang, Y., Zhou, X. and Zhou, J., Moses-based Chinese-Uyghur statistical machine translation systems. *Proc. 2010 IEEE Youth Conf. on Infor., Computing and Telecommunications*, Beijing, China, 186-189 (2010).
13. Caseli, H.d.M. and Nunes, I.A., Statistical machine translation: little changes big impacts. *Proc. 2009 Seventh Brazilian Symp. in Infor. and Human Language Technol.*, São Carlos, São Paulo, Brazil, 63-71 (2009).
14. Wibawa, A.P., Nafalski, A., Kadarisman, A.E. and Mahmudy, W.F., Indonesian-to-Javanese machine translation. *Inter. J. of Innovation, Manage. and Technol.*, 4, 451-454 (2013).
15. Saboor, A. and Khan, M.A., Lexical-semantic divergence in Urdu-to-English example based machine translation. *Proc. 2010 6th Inter. Conf. on Emerging Technologies (ICET)*, Islamabad, Pakistan, 316-320 (2010).
16. Quirk, C. and Menezes, A., Dependency treelet translation: the convergence of statistical and example-based machine-translation? *Machine Translation,* 20, 43-65 (2006).
17. Wedhawati, W., Nurlina, W.E.S., Setiyanto, E. and Sukesti, R., *Latest Structure of Javanese Language.* Yogyakarta: Kanisius (2006).
18. Quinn, G., Teaching Javanese respect usage to foreign learners. *Electronic J. of Foreign Language Teaching*, 8, 362-370 (2011).
19. Suwadji, S., Javanese language today. *Lokakarya Pengajaran Bahasa dan Sastra Jawa*, Yogyakarta, 55-61 (1996).
20. Subroto, D.E., Rahardjo, M.D. and Setiawan, B., Endangered krama and krama inggil varieties of the Javanese language. *Linguistik Indonesia,* 26, 89-96 (2008).
21. Irvine, J.T., Style *as Distinctiveness: the Culture and Ideology of Linguistic Differentiation*. In: Eckert, P. and Rickford, J.R. (Eds), Style and Sociolinguistic Variation, Cambridge University Press (2002).
22. Goebel, Z., Enregisterment and appropriation in Javanese-Indonesian bilingual talk. *Language in Society*, 36, 511-531 (2007).
23. Purwadi, P., Mahmudi M. and Setijaningrum, E., *Javanese Language Structure*. Yogyakarta: Media Abadi (2005).
24. Setiyanto, A.B., *Parama Satra: Javanese Language*. Yogyakarta: Panji Pustaka (2010).
25. Wibawa, A.P., Nafalski, A., Murray, N. and Mahmudy, W.F., Edit distance algorithm to increase storage efficiency of Javanese corpora. *Proc. Inter. Conf. on Computer, Electrical, and Systems Sciences, and Engng.*, Singapore, 1056-1060 (2012).
26. Wibawa, A.P., Nafalski, A. and Mahmudy, W.F., Javanese speech levels machine translation: improved parallel text alignment based on impossible pair limitation. *Proc. IEEE Inter. Conf. on Computational Intelligence and Cybernetics*, Yogyakarta (2013).

27. Wibawa, A.P. and Nafalski, A., Intelligent tutoring system: a proposed approach to Javanese language learning in Indonesia. *World Transactions on Engng. and Technol. Educ*., 8, **2**, 216-220 (2010).
28. Wibawa, S., Teaching approach for Javanese language learning. *Lokakarya Pengajaran Bahasa dan Satra Jawa*, 41-54 (1996).
29. Wibawa, S., Efforts to maintain and develop Javanese language politeness. *Proc. Inter. Seminar of Javanese Language*, Paramaribo, Suriname, 1-10 (2005).
30. Nugrahani, F., Reactualisation of Javanese language and literature learning in multicultural era. *Varia Pendidikan*, 20, 70-80 (2008).